
Directed-Tokens: A Robust Multi-Modality Alignment Approach to Large Language-Vision Models

Anonymous Author(s)

Affiliation

Address

email

1 A Benchmarks

2 Following the standard benchmarks of LLaVa v1.5, we evaluate our models on two sets of bench-
3 marks, i.e., Academic Task-oriented Benchmarks and Instruction-Following LMM Benchmarks.
4 For the academic task-oriented benchmarks, we adopt five different benchmarks, including Visual
5 Question Answering V2 (VQAv2) [1], Question Answering on Image Scene Graphs (GQA) [3],
6 Answer Visual Questions from People Who Are Blind (VizWiz) [2], Science Question Answer-
7 ing (SciQA-IMG) [11], and Visual Reasoning based on Text in Images (TextVQA) [12]. While
8 VQAv2 and GQA focus on evaluating the visual understanding based on the open-ended short an-
9 swers, VizWiz evaluates the generalization of the model based on the visual questions raised by
10 impaired people. On the other hand, SciQA-IMG benchmarks will measure the performance of the
11 LMM on scientific questions via multiple-choice questions. The TextVQA benchmark evaluates
12 the capability of models in reading and reasoning about text in images. Meanwhile, we use six
13 Instruction-Following LMM Benchmarks to evaluate our proposed approach, including Polling-
14 based Object Probing Evaluation for Object Hallucination (POPE) [7], Multimodal LLMs with
15 Generative Comprehension Benchmark (SEED-Bench) [5], Comprehensive Evaluation Benchmark
16 of LMM (MME) [13], LLaVA Benchmark in the Wild (LLaVA-Wild) [9], Integrated Capability
17 Benchmark (MM-Vet) [14], and Massive Multi-discipline Multimodal Understanding and Reason-
18 ing Benchmark (MMM-U-Val) [15]. While the POPE benchmark evaluates the hallucination of
19 the model based on the tree subset, i.e., random (rand), common (pop adv), and adversarial (adv),
20 SEED-Bench measures the generative comprehension of the LMM on both images and videos with
21 multiple-choice questions. For the video evaluation, we adopt the middle frame of the video as a vi-
22 sual input. The MME perception benchmark evaluates visual understanding of the LMM via binary
23 (yes/no) questions. Meanwhile, the LLaVA-Wild and MM-Vet benchmarks measure the capabili-
24 ties of the LMM in engaging in visual conversations. The MMMU benchmark evaluates LMMs on
25 massive multi-discipline tasks demanding high-level subject knowledge and reasoning.

26 B Additional Ablation Study

27 **Effectiveness of Data Size.** Table 1 presents a comparative analysis of LLaVA-7B and Direct-
28 LLaVA-7B performance across varying data sizes during pre-training and fine-tuning, evaluated
29 on a range of LLM benchmarks. We use the data training size of LLaVA v1 [9] and LLaVA
30 v1.5 [8] for our experiments. Overall, both models demonstrate enhanced outcomes with larger
31 fine-tuning datasets, with Direct-LLaVA consistently surpassing LLaVA. Notably, performance on
32 instruction-following benchmarks, such as POPE, MME, and SEED-Bench, improves substantially
33 as fine-tuning data increases. Specifically, the F1 score on the POPE benchmark rises by approxi-
34 mately 12.5% across different configurations. For the MME benchmark, the scores of LLaVA and
35 Direct-LLaVA increase markedly from 809.6 and 1102.1 to 1510.7 and 1524.9, respectively. The

accuracy gain on SEED-Bench is also significant, with larger datasets nearly doubling the models’ performance. Furthermore, Direct-LLaVA-7B consistently outperforms LLaVA across both academic-oriented and instruction-following benchmarks. For instance, on academic benchmarks, Direct-LLaVA-7B achieves a 7.5% and 5% higher accuracy in ScienceQA and TextVQA, respectively. These results underscore the robustness of our proposed method.

Table 1: Effectiveness of Pre-training and Fine-tuning Data Size.

Method	Data Size		SciQA IMG	Text VQA	POPE			MME	SEED-Bench		
	Pretrain	Finetune			rand	pop	adv		all	img	vid
LLaVA-7B	595K	158K	46.9	26.1	76.3	72.2	70.1	809.6	33.5	37.0	23.8
Direct-LLaVA-7B	595K	158K	54.6	29.3	79.5	76.2	74.3	1102.1	36.5	40.3	27.7
LLaVA-v1.5-7B	558K	665K	66.8	58.2	87.3	86.1	84.2	1510.7	58.6	66.1	37.3
Direct-LLaVA-7B	558K	665K	74.3	63.2	88.8	88.9	86.0	1524.9	63.3	69.7	38.8

Scalability to Larger Data and Benchmarks. To illustrate our scalability to larger data and other benchmarks, we conduct experiments on LLaVA-OneVision data [6] with LLaVA and Qwen2.5-0.5B. We report our results on MME, MMMU, SeedBench-IMG, AI2D [4], and MMBench [10]. As shown in Table 2, when data is scaling up, our proposed approach still maintains its effectiveness and significantly improves the performance of LMM on various benchmarks.

Table 2: Effectiveness of Direct-LLaVA on Large Dataset.

	% Samples for Pretext	MME	MMMU	SeedBench-IMG	AI2D	MMBench
LLaVA	-	1238	31.4	65.5	57.1	52.1
Direct-LLaVA	50%	1351	32.9	67.5	62.6	55.4
Direct-LLaVA	100%	1494	34.5	68.4	69.4	58.7

Effectiveness of Data In Pretext. To understand the effectiveness of our proposed approach on the ratio of data used, we conducted an experiment using only 50% of data for our pretext tasks. As shown in Table 2, our approach can improve the performance of the LMMs. The results have further confirmed the effectiveness of our proposed learning approach.

Visualization of Shuffle Predictions. We provide the real-world images to illustrate our effectiveness of the shuffling learning mechanism. As shown in Figure 1, for Direct-LLaVA with `cls` (no text in reconstruction), the LMM predicts the image order well but shows noticeable differences from the originals. In contrast, Direct-LLaVA with `drt` (text included) better aligns with the original images since information of language and visuals are well captured in `drt` token. To highlight the impact of text in reconstruction, we altered the description. Although image patches became inconsistent, the images were adapted to match the semantic meaning of the text.

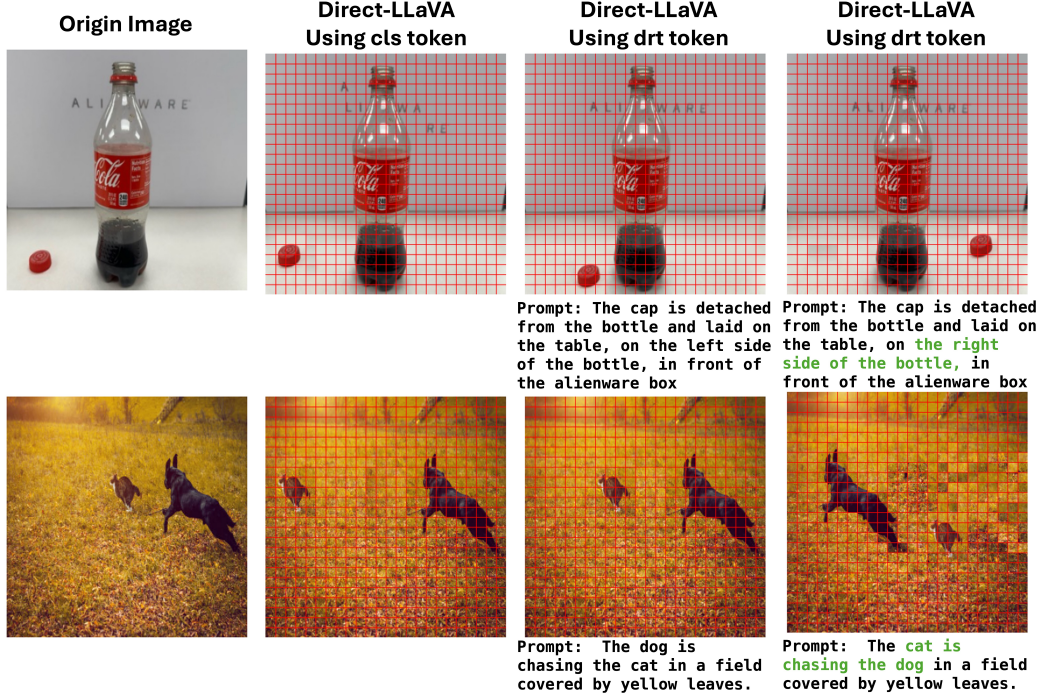


Figure 1: Additional Visualization (Better in $2\times$ zoom for details).

57 C Discussion of Limitations

58 Our paper has adopted a specific set of hyper-parameters and learning methods to support our hypothesis. However, our work could contain several limitations. Our work investigated the effectiveness of our proposed learning tasks and losses in improving the LMM’s performance. Thus, 60 the investigation of balance weights among learning objectives has not been fully exploited, and 62 we leave this experiment as our future work. Due to computation limitations, our experiments are limited to the standard language model size (i.e., Vicuna 13B, Vicuna 7B, LLaMA3 8B, and Qwen 64 7B) and data scale (LLaVA v1.5). Nevertheless, we hypothesize that our proposed approaches can generalize to larger-scale language models and data settings due to their fundamental theories. 65

66 References

- 67 [1] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- 70 [2] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- 73 [3] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- 76 [4] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images, 2016.
- 78 [5] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- 80 [6] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

- 82 [7] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in
83 large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- 84 [8] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In
85 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages
86 26296–26306, 2024.
- 87 [9] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information
88 processing systems*, 36, 2024.
- 89 [10] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen,
90 and D. Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*,
91 2023.
- 92 [11] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan.
93 Learn to explain: Multimodal reasoning via thought chains for science question answering.
94 *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- 95 [12] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach.
96 Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer
97 vision and pattern recognition*, pages 8317–8326, 2019.
- 98 [13] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large
99 language models. *arXiv preprint arXiv:2306.13549*, 2023.
- 100 [14] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating
101 large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- 102 [15] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun,
103 C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su,
104 and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning
105 benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision
106 and Pattern Recognition*, pages 9556–9567, 2024.